small number of tertiary constraints. Important aspects of the invention include its utilization of side chain center of mass representations and the very small number of tertiary constraints required to assemble moderate resolution folds. For a representative set of al types of single domain proteins (all-α, all-β, α/β motifs), the required number of constraints is about N/7, with N the number of residues in the protein. Furthermore, due to a new, rapid treatment of side chain burial, the invention is applicable to multi-domain proteins.

The invention also provides a relatively simple and reliable protocol of detecting a proper fold from less frequently generated misfolded structures. These misfolded structures are almost exclusively the topological mirror images of the proper fold. In all cases examined to date, the native-like structure always has a lower conformational energy. This and the small number of required tertiary constraints suggest that the invention's underlying force field captures a number of the essential aspects of protein interactions. At the same time, the model is simpler and computationally more efficient than previously employed lattice models.[6,39] Due to a much lower computational cost (at least one order of magnitude), it is possible to assemble larger structures, including the 247-residue Atim domain.

Finally, using the invention it is also possible to generate at least low resolution folds using only a small set of probable side chain contacts[35] (predicted via correlated mutations analysis[43]) and somewhat more elaborate potentials describing short-range interactions (derived from geometrical analysis of sequentially similar protein fragments). Several example structures such as myohemerythrin (1hmd) and the complex β-type motif immunoglobulin fold (1fna), have been assembled.

Example 2

This example also describes the generation and refinement of protein molecular models. Threading-based target-template alignments were obtained from one standard threading method;[15] but in principle, any could be used.[26] The modeling technique employed was SICHO, which employs a very simple, and computationally very efficient, yet quite accurate, representation of protein structure and dynamics.[17,19] For the purpose of the this application, the model was refined by incorporating evolutionary information into the interaction scheme. Starting from an initial conformation of the model lattice chain that approximately followed the threading template, a Monte Carlo annealing procedure found a conformation that maintained some (but not all) features of the original template and at the same time optimized packing and intra-protein interactions, as defined by the reduced model of the probe protein. This could be also visualized as a folding simulation in a soft tube built around the threading template.

Here, the method was applied to 12 target/template protein pairs that produce various quality models. The parameters of the lattice model force field (more precisely, the balance between the intrinsic force field and the template-related biases) were adjusted by a trial and error method for three of the 12 target/template protein pairs. The obtained parameters were subsequently used in the other 9 simulations. As will become apparent after analysis of the simulation results, the obtained models for the three proteins used for timing the potential were among the best. This may suggest that the method was strongly tuned to these three examples. This was not the case. First, the three proteins belonged to completely different structural classes, so any tuning was rather general, *i.e.*, applicable to the majority of single domain proteins. Second, when the tuning procedure was performed on just a single case (the plastocyanin/azurin pair), almost the same results are obtained, suggesting that the optimal balance between the template-related soft restraints and

the intrinsic force field of the model was similar for various proteins. Finally, the poorer results obtained for most of the remaining nine test proteins were simply due to the poor quality of the initial threading models.

The remainder of this example can be outlined as follows. In Methods, the reduced lattice protein model is described, with the protein representation, the model of stochastic dynamics, the interaction scheme, and the template related biases and restraints being discussed. Then, in the Results section, the molecular models obtained from Monte Carlo simulated annealing and subsequent refinement procedures are compared with the initial crude, threading-based models. In the Discussion, the improved models are analyzed to identify typical underlying structural rearrangements.. Certain technical details are found in the Appendix.

Methods

Lattice model

The reduced modeling of protein structure and dynamics usually employs an alpha carbon main chain representation.[18,25] Side chains are either completely neglected or treated at various levels of simplification. The choice of the alpha carbon representation is mostly motivated by the high level of geometric regularity of the main chains in folded proteins.[25] On the other hand, the packing and interactions between the side chains are perhaps much more sequence specific than are those of the main chain. The latter are very similar in all proteins.

As demonstrated in Example 1, SICHO is a useful protein-modeling tool, as it incorporates many protein-like features, including local conformational propensities and the characteristic packing regularities of protein side chains. A major advantage of SICHO is that the entire conformational space of quite large proteins can be efficiently sampled. For example, with the help of a properly designed force field, loose knowledge of the secondary structure and a few long-range side chain contacts (about N/7, where N is the number of residues), which